Comprehensive variant detection in a human genome with highly accurate long reads

Kirrenjit Kaur, William J. Rowell, Aaron M. Wenger

Pacific Biosciences, Menlo Park, CA, USA

Introduction

- Long-read sequencing has been applied successfully to assemble genomes and detect structural variants.
- Long reads can be unambiguously mapped to more of the genome than short reads of comparable accuracy.
- It has been difficult to call small variants from long reads due to higher read error rates.
- PacBio Circular Consensus Sequencing (CCS) produces **>99% accurate, 10 kb - 20 kb HiFi reads** that enable detection of small (SNVs and <50 bp indels) and large (≥50 bp) variants.

Methods

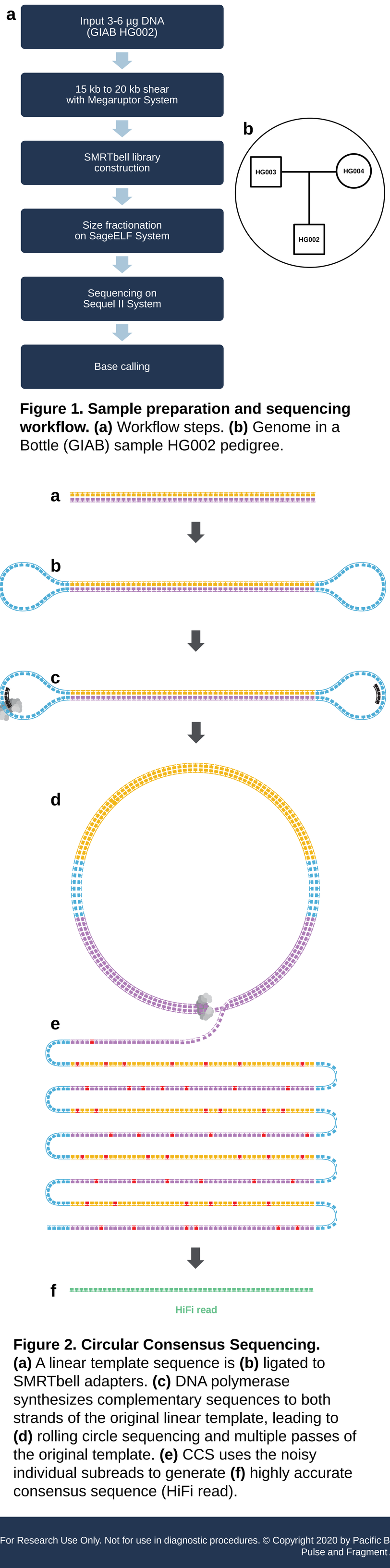

Figure 1. Sample preparation and sequencing workflow. (a) Workflow steps. **(b)** Genome in a Bottle (GIAB) sample HG002 pedigree.

Figure 2. Circular Consensus Sequencing. (a) A linear template sequence is **(b)** ligated to SMRTbell adapters. **(c)** DNA polymerase synthesizes complementary sequences to both strands of the original linear template, leading to **(d)** rolling circle sequencing and multiple passes of the original template. **(e)** CCS uses the noisy individual subreads to generate **(f)** highly accurate consensus sequence (HiFi read).

Analysis

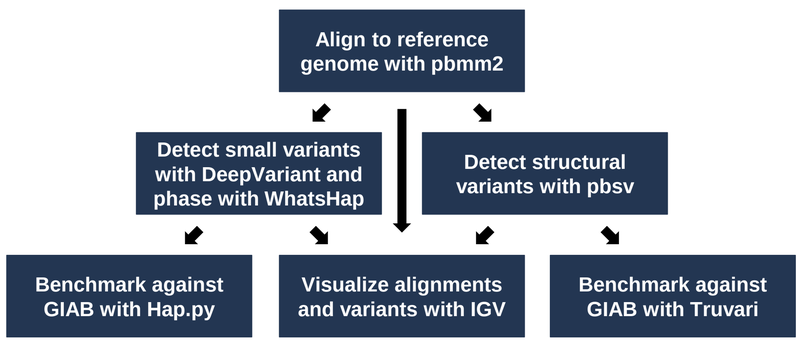

Figure 3. Bioinformatics workflow for read mapping and variant detection. Approximately 18-fold coverage (two SMRT Cells 8M) of highly accurate (average 99.8%) 12.9 kb reads were mapped to the hg19 reference with pbmm2. Single nucleotide variants (SNVs) and small indels (<50 bp) were detected using Google DeepVariant v0.10.0 with a model trained for CCS. Structural variants (SVs) were detected with pbsv v2.3.0. Variant calls were evaluated against GIAB benchmarks.

Benchmarking

Detection	Type	Recall	Precision
DeepVariant	SNV	99.6%	99.9%
	indel (<50 bp)	95.8%	96.8%
pbsv	SV (≥50 bp)	97.7%	95.3%

Table 1. Variant detection compared against Genome in a Bottle benchmarks. SNVs and indels <50 bp were compared against the GIAB HG002 v4.1 small variant benchmark using Hap.py. SVs were compared against the GIAB HG002 v0.6 SV benchmark using Truvari.

Median block length	25 kb
Mean block length	99 kb
Max block length	2.3 Mb
Sum of phased bases	2.0 Gb
Phase block N50	177 kb
Phase switch error rate	0.35 %

Table 2. Small variant phasing metrics. Small variants detected by DeepVariant were phased using HiFi reads. Blocks size metrics were generated by WhatsHap. Switch errors were measured against GIAB 10x trio-phased VCF.

Comprehensive Variant Calling in Difficult Regions

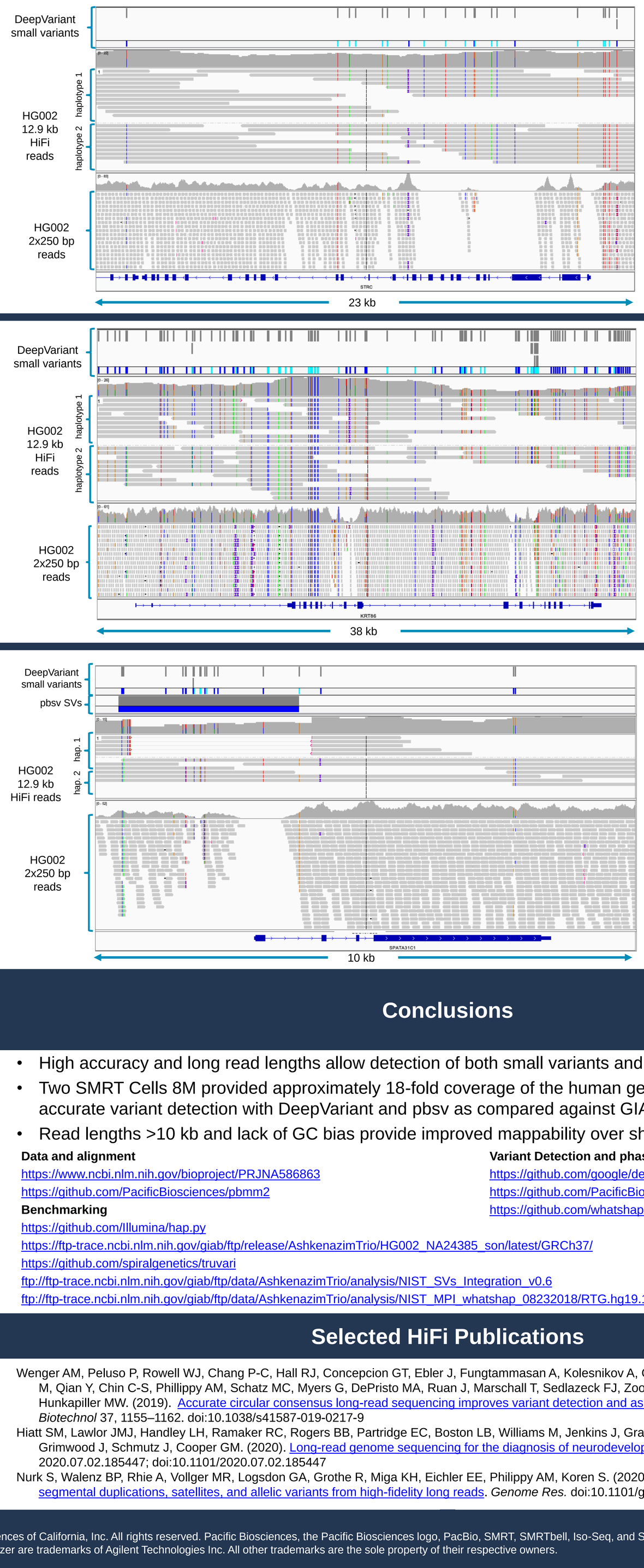

Figure 4. *STRC* locus. Segmental duplications within this locus prevent the unambiguous alignment of short reads. HiFi reads can be aligned not only to all exons at this locus, but intronic sequence as well, allowing for small variants to be called by DeepVariant across the entire locus.

Figure 5. *KRT86* locus. The length and accuracy of HiFi reads allow small variant calling even in reference-divergent regions within segmental duplications.

Figure 6. *SPATA31C1* locus. Longer read lengths allow pbsv to detect a heterozygous 3.4 kb deletion covering the first exon.

Conclusions

- High accuracy and long read lengths allow detection of both small variants and structural variants.
- Two SMRT Cells 8M provided approximately 18-fold coverage of the human genome, which achieves highly accurate variant detection with DeepVariant and pbsv as compared against GIAB benchmarks.
- Read lengths >10 kb and lack of GC bias provide improved mappability over short-read sequencing.

Data and alignment

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA586863

https://github.com/PacificBiosciences/pbmm2

Benchmarking

https://github.com/Illumina/hap.py

https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh37/

https://github.com/spiralgenetics/truvari

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_MPI_whatshap_08232018/RTG.hg19.10x.trio-whatshap.vcf.gz

Variant Detection and phasing

https://github.com/google/deepvariant

https://github.com/PacificBiosciences/pbsv

https://github.com/whatshap/whatshap/

Selected HiFi Publications

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR & Hunkapiller MW. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 37, 1155–1162. doi:10.1038/s41587-019-0217-9

Hiatt SM, Lawlor JMJ, Handley LH, Ramaker RC, Rogers BB, Partridge EC, Boston LB, Williams M, Jenkins J, Gray DE, Holt JM, Bowling KM, Bebin EM, Grimwood J, Schmutz J, Cooper GM. (2020). Long-read genome sequencing for the diagnosis of neurodevelopmental disorders. *bioRxiv* 2020.07.02.185447; doi:10.1101/2020.07.02.185447

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Philippy AM, Koren S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* doi:10.1101/gr.263566.120